THE EFFECT AND ADJUSTMENT OF COMPLEX SURVEYS ON CHI-SQUARED GOODNESS OF FIT TESTS.SOME MONTECARLO EVIDENCE.

Victor Aguirre-Torres, ITAM, Alejandro Rios-Curiel
Victor Aguirre-Torres, Departamento de Estadistica y Actuaria, ITAM, Rio Hondo No. 1, Mexico D.F. 01000, MEXICO

Key Words: Sampling complexity, Pearson type tests, Wald type tests, Average eigenvalue correction, Satterthwaites's correction.

1. INTRODUCTION

Most of the inferential results are based on the assumption that the user has a "random" sample, by this it is usually understood that the observations are a realization from a set of independent identically distributed random variables. However most of the time this is not true mainly for two reasons: one, the data are not obtained by means of a probabilistic sampling scheme from the population, the data are just gathered as they becomes available or in the best of the cases using some kind of control variables and quota sampling.; and second, even if a probabilistic scheme is used, the sample design is complex in the sense that it was not simple random sampling with replacement, but instead some sort of stratification or clustering or a combination of both was required. For an excellent discussion about the kind of considerations that should be made in the first situation see Hahn and Meeker (1993) and a related comment in Aguirre (1994). For the second problem there is a book about the topic in Skinner et al.(1989). In this paper we consider the problem of evaluating the effect of sampling complexity on Pearson's Chi-square and other alternative tests for goodness of fit for proportions. Work on this problem can be found in Shuster and Downing (1976), Rao and Scott (1974), Fellegi (1980), Holt et al. (1980), Rao and Scott (1981), and Thomas and Rao (1987). Out of this work come up several adjustments to Pearson's test, namely: Wald type tests, average eigenvalue correction and Satterthwaite type correction. There is a more recent and general resampling approach given in Sitter (1992), but it was not pursued in this study.

The paper is organized as follows: section 2 reviews some general large sample results regarding Pearson's and Wald type tests for this problem, special consideration is made for stratified and two stage cluster sampling. Section 3 gives some easy to compute adjustments to Pearson's test that use the large sample results of section 2 in terms of a function of the eigenvalues of the large sample distribution. Section 4 presents the results of comparing the various statistics under different populations and sampling schemes, significance levels are compared.

2. LARGE SAMPLE RESULTS.

Consider the problem of testing that the population proportions of a categorical variable X with k possible values has some predetermined values, against the alternative that for at least one category they are different. That is, if we let

$$P(X = i) = p_i$$

then we want to test

H_0: $p_i = p_{0i}$ for every i = 1,...,k
H_1: p_i not equal to p_{0i} for some i

Pearson's chi-square test is given by

$$X^2_P = n \sum_{i=1}^{k} (\hat{p}_i - p_{0i})^2 / p_{0i}$$

where $\hat{p}_i$ is a consistent estimator of p_i under the sampling scheme used.

It is easy to show, Serfling (1980), that for simple random sampling with replacement (SRSWR) X^2_P has an asymptotic chi-square distribution with k-1 degrees of freedom ($\chi^2(k-1)$). A more general result given in Johnson and Kotz (1970) states that under H_0 X^2_P is distributed asymptotically as

$$\sum_{i=1}^{k-1} \lambda_{0i} Z_i^2$$

where the Z_i^2 are independent chi-square random variables with one degree of freedom, and the λ_{0i} 's are the eigenvalues of the matrix of design effects defined by

$$D_0 = P_0^{-1} V_0$$

if we let $\mathbf{p}_0$ be the column vector of probabilities under the null hypothesis then the matrix P_0 is defined by

$$P_0 = \text{diagonal}(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0^t \qquad (1)$$

and the matrix V_0/n is the variance covariance of $\hat{p}$ under H_0, under simple random sampling P_0 and V_0 are the same and hence the eigenvalues are all equal to one and one gets the previous result, but under a complex sampling plan that is no longer the case and the large sample distribution is not a $\chi^2(k-1)$.

It may be shown that Pearson's test is a quadratic form of the vector $\hat{p} - \mathbf{p}_0$ with P_0 in the middle, that is why the $\chi^2(k-1)$ is right for SRSWR and wrong under a complex sampling. One way to get around this problem is to use the correct matrix for the quadratic, this is the Wald type test statistic. In general the test statistic may be written as

$$X^2_W = n(\hat{p} - \mathbf{p}_0)^t V^{-1} (\hat{p} - \mathbf{p}_0)$$

with V/n the asymptotic variance covariance matrix of $\hat{p}$ under the corresponding sampling plan and under the null hypothesis. The matrix V depends heavily on the sampling plan being used, to see this we will consider two cases: stratified random sampling, and cluster sampling.

Stratified random sampling.

Consider the following notation:

L= number of strata
N_h= number of elements in stratum h
N= number of elements in the population
$W_h = N_h/N$
n_h = sample size stratum h = nW_h
y_{jhi}= 1 if the j-th element of the sample from h-th stratum belongs to the i-th category, and zero otherwise
y_{jh}= vector with entries y_{jhi}
$\mathbf{p}_h$= vector of strata proportions of each category
$\mathbf{p}$= vector of population proportions of each category
P= matrix defined in (1) with $\mathbf{p}$
$\hat{p}_h$= sample estimate of $\mathbf{p}_h$

$$\hat{p} = \sum_{h=1}^{L} W_h \hat{p}_h$$

for stratified random sampling with proportional allocation and SRSWR within each stratum V becomes, see Rao and Scott (1981):

$$V = P - \sum_{h=1}^{L} W_h(\mathbf{p}_h - \mathbf{p})(\mathbf{p}_h - \mathbf{p})^t$$

if we let

$$\hat{V}_h = (n_h(n_h - 1))^{-1} \sum_{j=1}^{n_h} (y_{jh} - \hat{p}_h)(y_{jh} - \hat{p}_h)'$$

then a consistent estimator of V is

$$\hat{V} = n \sum_{h=1}^{L} W_h^2 \hat{V}_h$$

in the Monte Carlo study we are going to consider two different Wald tests, X^2_{WA} with V in the Middle and X^2_{WB} with $\hat{V}$. Of course X^2_{WA} can not be computed from the sample but we wanted to observe the possible performance of Wald's test when there is no sampling variation in the estimation of V.

Two stage cluster sampling

We consider two stage cluster sampling with first stage selection made with probability proportional to size of the cluster, and the second stage SRSWR. For this purpose let

R= Number of clusters in the population
N= Total number of elements in the population
r = Number of clusters in the sample
M_i= Number of elements within the i-th cluster
m_l = number of elements in the sample within the l-th cluster
n = Total sample size = r m
W_i= probability of selection of the i-th cluster = M_i /N

$\mathbf{p}_i$ = vector of cluster proportions for each category

$\hat{p}_j$ = the vector of within cluster sampling proportions, j= 1, 2, ..., r

$$\hat{p} = \sum_{j=1}^{r} \hat{p}_j$$

p= vector of population proportions of each category

P= matrix defined in (1) with **p**

then Rao and Scott (1981) mention that

$$V = P + (m-1) \sum_{i=1}^{R} W_i (p_i - p)(p_i - p)^t$$

an unbiased consistent estimator of V is

$$\hat{V} = m \sum_{j=1}^{r} (\hat{p}_j - \hat{p})(\hat{p}_j - \hat{p})^t$$

As before, for the Monte Carlo study we considered two different Wald Type tests X^2_{WA} and X^2_{WB}.

It is clear from these two examples that the impact of sampling complexity on the test statistic is through the difference between V and P.

3. ADJUSTMENTS VIA EIGENVALUES.

Notice also that Wald type B test statistic require the estimation of the whole V^{-1}, this is usually done by estimating V first, as shown above, and then inverting the estimator. This procedure may be numerically unstable as will be shown in the simulation results. An alternative to this task is shown in this section, instead of estimating a matrix, the adjustments depend on some functions of the eigenvalues which in turn are simple functions of the elements of V.

First consider the average eigenvalue adjustment (AVE), it consists of dividing Pearson's test by the average of the first k-1 eigenvalues. The rational is as follows, under sampling complexity

$$X^2_P \xrightarrow{L} \sum_{i=1}^{k-1} \lambda_{0i} Z_i^2$$

if we let

$$\lambda_{0\cdot} = (k-1)^{-1} \sum_{i=1}^{k-1} \lambda_{0i}$$

then

$$X^2_P / \lambda_{0\cdot} \xrightarrow{L} \sum_{i=1}^{k-1} (\lambda_{0i} / \lambda_{0\cdot}) Z_i^2$$

if the eigenvalues are not too far apart then the limiting distribution would be approximately $\chi^2(k-1)$, see Rao and Scott (1981). The nice thing about this correction is that it is very simple to compute an use, because

$$\lambda_{0\cdot} = \mathrm{tr}(P^{-1} V) = \sum_{i=1}^{k} v_{ii} / (p_i[k-1])$$

where v_{ii} are the diagonal elements of V, therefore a consistent estimator $\lambda_{\cdot}$ can be obtained by replacing the sample counterparts into the above formula. And from a direct application of Slutsky's theorem we get

$$X^2_{AVE} = X^2_P / \lambda_{\cdot} \xrightarrow{L} \sum_{i=1}^{k-1} (\lambda_{0i} / \lambda_{0\cdot}) Z_i^2$$

Applications of this idea may be found in Holt, Scott, and Ewings (1980). It is important to notice that it is an empirical procedure that has given good results in practice.

We now consider Satterthwaite (1946) correction, it is based on the observation that the limiting distribution of X^2_P is a weighted sum of mean squares with one degree of freedom each, from here the correction consists of approximating the distribution of the random variable with a Chi-square distribution where the number of degrees of freedom are estimated. In this paper we adopt the form of Satterthwaite's' correction (SAT) for X^2_P given in Rao and Scott (1981)

$$X^2_{SAT} = X^2_P / (\lambda_{\cdot} (1+a^2))$$

where

$$a^2 = \sum_{i=1}^{k-1} (\lambda_i - \lambda_{\cdot})^2 / [(k-1)\lambda_{\cdot}^2]$$

and

$$\sum_{i=1}^{k-1}\lambda_i^{\,2} = \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}^2 / (\hat{p}_i \hat{p}_j)$$

the v_{ij}'s are the elements of V. To check for significance X^2_{SAT} is compared with a critical value from a Chi-square distribution with $(k-1)/(1+a^2)$ degrees of freedom.

Notice the computational advantage of X^2_{AVE} and X^2_{SAT} , they are based on the usual X^2_P and just require the estimation of some simple functions of the elements of V.

4. MONTE CARLO RESULTS.

For the Monte Carlo study we considered stratified and cluster sampling, two population sizes (500 and 1000), and two sampling fractions (.1 and .2). In cluster sampling we considered two cluster sizes (20 and variable). The program that performed the simulations was programmed using Lotus-123. The populations were generated with the true $\mathbf{p}_0$ equal to (.1, .4, .4, .1). To form the strata, the population was ordered first and then broken down in a proportion .4, .2, .4. For cluster sampling with five hundred elements, the population was just broken down in 25 clusters. For the population of size one thousand, there were five clusters of size one hundred, five of size fifty, and ten of size twenty five,Tables 1a and 1b show the specific scenarios for the simulation as well as the proportion of test statistics that exceeded the 5% critical point after 200 replications.The first for correspond to stratified random sampling, while the rest are cluster sampling. We also analyzed the 10%, 2.5%, and 1% tails, the results were similar.

It is easy to spot from table 1 that the best test was X^2_{WA} , while the worst was X^2_{WB} . To see the situation, figure 1 shows the empirical sizes of the tests stratified with respect to sampling plan. Pearson's test is conservative in stratified sampling, and becomes much more liberal in cluster sampling. Type A Wald test is OK under both sampling plans. Type B Wald test, the one that would be computed from the sample data, is invalid under both sampling schemes, it is worst under cluster sampling. Pearson's average correction is above 5% under both sampling plans, but not too far away, particularly in cluster sampling. Pearson's Satterthwaite's correction behaves similarly to the average correction under stratified sampling, but is much better under cluster sampling, in fact it is quite similar to X^2_{WA} .

Pop Size	Sampling Fraction	Cluster Size	X^2_P	X^2_{WA}	X^2_{WB}
500	.1 Strat	n/a	3.5	7.0	15.5
500	.2	n/a	2.0	4.0	7.0
1000	.1	n/a	1.5	6.0	11.0
1000	.2	n/a	5.5	7.5	16.5
500	.1 Clust	20	10.5	5.0	19.5
500	.1	20	15.5	6.0	48.0
500	.2	20	17.5	4.0	23.0
1000	.1	variable	8.5	6.0	19.5
1000	.1	variable	6.5	4.0	39.0
1000	.1	variable	6.0	4.5	12.0
1000	.2	variable	14.5	5.5	27.0
1000	.2	variable	20.5	6.0	47.5
1000	.2	variable	7.0	3.5	7.5

Table 1.a. Percent of test statistics exceeding a 5% critical point. Tests: Pearson, WA, WB.

Pop Size	Sampling Fraction	Cluster Size	X^2_{AVE}	X^2_{SAT}
500	.1 Strat	n/a	15.0	17.0
500	.2	n/a	6.5	7.0
1000	.1	n/a	6.5	6.5
1000	.2	n/a	15.5	16.5
500	.1 Clust	20	7.0	3.5
500	.1	20	8.5	5.0
500	.2	20	5.5	3.5
1000	.1	variable	7.0	6.5
1000	.1	variable	6.0	6.0
1000	.1	variable	9.0	6.0
1000	.2	variable	7.5	5.5
1000	.2	variable	10.0	7.0
1000	.2	variable	4.5	3.5

Table 1.b. Percent of test statistics exceeding a 5% critical point. Tests: AVE, SAT.

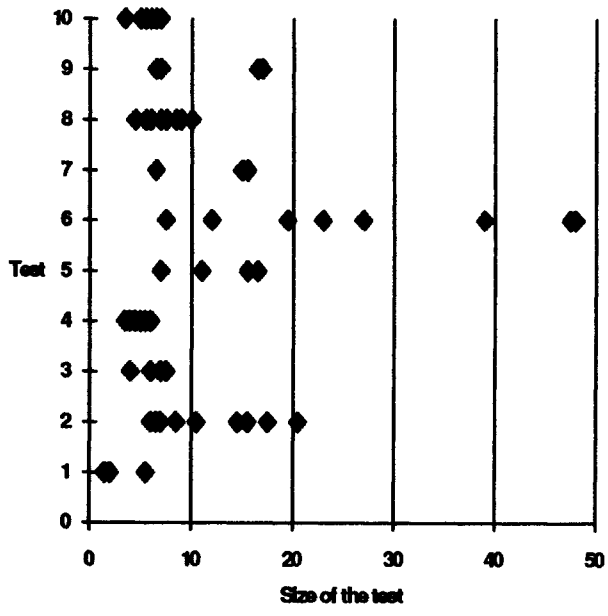

Figure 1. Comparison of tests. Nominal size 5%. Sampling: stratified (odd number), cluster (even number). Pearson's Test: 1 and 2, Wald A Test: 3 and 4, Wald B Test: 5 and 6,Average correction: 7 and 8, Satterthwaite' s correction: 9 and 10.

To learn a little more about the performance of the tests, table 2 gives the correlation coefficients of the empirical sizes of the tests.

	X^2_P	X^2_{WA}	X^2_{WB}	X^2_{AVE}
X^2_{WA}	-.023			
X^2_{WB}	.732	.167		
X^2_{AVE}	-.135	.822	.036	
X^2_{SAT}	-.396	.745	-.164	.926

Table 2. Correlation between empirical sizes of the tests.

From that table one can see that:

a) There is no correlation between Pearson's test and X^2_{WB} , which is bad

b) Pearson's test and X^2_{WB} are correlated, something bad too

c) X^2_{AVE} and X^2_{SAT} are correlated with X^2_{WA} , something favorable

d) X^2_{AVE} and X^2_{SAT} are correlated, as expected

Figure 2 shows the association between the sizes of XWA and XAVE, it is interesting to see that they behaved similarly, although the impact of sampling complexity on XAVE was stronger.

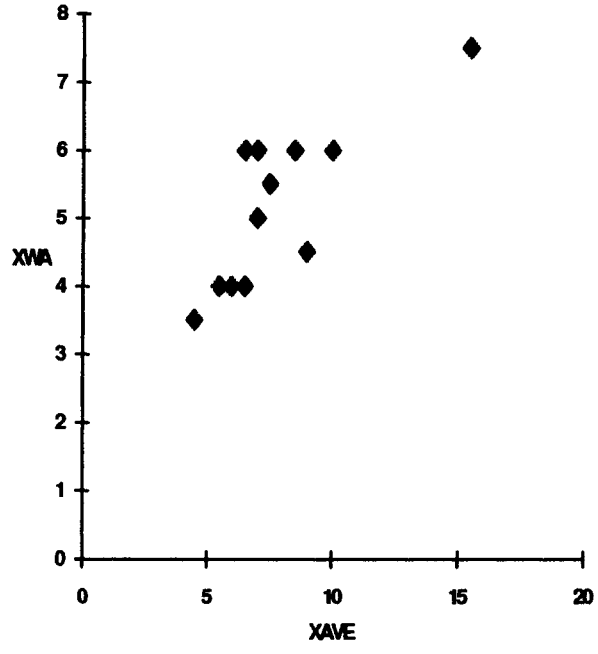

Figure 2. Dispersion diagram of empirical sizes for tests XWA and XAVE.

5. SUMMARY AND CONCLUSIONS.

The paper shows the importance of sampling complexity in the performance of goodness of fit tests for proportions. Pearson's test was conservative under stratified sampling and liberal under cluster sampling. Wald's test had by far the worst performance of all tests, this was caused by the estimation of inverse of the variance covariance matrix of the vector of estimated proportions, a better estimation of this inverse would result in an improved test that may perform well under both kinds of sampling. Eigenvalue corrections behaved similarly but Satterthwaite's test performed much better in cluster sampling.

REFERENCES

Aguirre, V. (1994). Comment on "Assumptions for Statistical Inference" by Hahn, G. J., and Meeker, W. Q. (1993). *The American Statistician* 48, 60.

Fellegi, I. P. (1980). "Approximate tests for independence and goodness of fit based on stratified multistage samples". *Journal of the American Statistical Association* 75, 261-268.

Hahn, G. J., and Meeker, W, Q. (1993). "Assumptions for statistical inference". *The American Statistician* 47, 1-11.

Holt, D., Scott, A. J., and Ewings, P. O. (1980). "Chi-squared tests with survey data". *Journal of the Royal Statistical Society*, series A 143, 302-320.

Johnson, W. D., and Kotz, S. (1970). *"Continuous Univariate Distributions"*, Hougton Mifflin, Boston.

Rao, J. N. K., and Scott, A. J. (1979). "Chi-squared tests for analysis of categorical data from complex surveys". *Proceedings of the American Statistical Association*, Section on survey Research Methods, 58-66.

Rao, J. N. K., and Scott, A. J. (1981). "The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two way tables". *Journal of the American Statistical Association* 76, 221-230.

Satterthwaite, F. E. (1946). "An approximate distribution of estimates of variance components". *Biometrics Bulletin* 2, 110-114.

Serfling, R. (1980). *"Approximation theorems of mathematical statistics"*. Wiley and sons, New York.

Shuster, J. J., and Downing, D. J. (1976). "Two way contingency tables for complex sampling schemes". *Biometrika* 63. 271-276.

Sitter, R. R. (1992). "A resampling procedure for complex survey data". *Journal of the American Statistical Association* 87, 755-765.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *"Analysis of complex surveys"*. Wiley and sons, New York.

Thomas, D. R., and Rao, J. N. K. (1987). "Small sample comparisons of level and power for simple goodness of fit statistics under cluster sampling". *Journal of the American Statistical Association* **82**, 630-636.